Charges Fit to Electrostatic Potentials. II. Can Atomic Charges Be Unambiguously Fit to Electrostatic Potentials?

MICHELLE MILLER FRANCL,* CHRISTINA CAREY, and LISA EMILY CHIRLIAN

Department of Chemistry, Bryn Mawr College, 101 N. Merion Avenue, Bryn Mawr, Pennsylvania 19010-2899

DAVID M. GANGE*

American Cyanamid Company, Agricultural Research Division, P.O. Box 400, Princeton, New Jersey 08543-0400

Received 25 October 1994; accepted 10 May 1995

ABSTRACT

The present work examines the conditioning of the least-squares matrix for obtaining potential derived charges and presents a modification of the CHELP method for fitting atomic charges to electrostatic potentials. Results from singular value decompositions (SVDs) of the least-squares matrices show that, in general, the least-squares matrix for this fitting problem will be rank deficient. Thus, statistically valid charges cannot be assigned to all the atoms in a given molecule. We find also that, contrary to popular notions, increasing the point density of the fit has little or no influence on the rank of the problem. Improvement in the rank can best be achieved by selecting points closer to the molecular surface. Basis set has, as expected, no effect on the number of charges that can be assigned. Finally, a well-defined, computationally efficient algorithm (CHELP-SVD) is presented for determining the rank of the least-squares matrix in potential-derived charge fitting schemes, selecting the appropriate subset of atoms to which charges can be assigned based on that rank estimate, and then refitting the selected set of charges.

* Authors to whom all correspondence should be addressed.

CCC 0192-8651 / 96 / 030367-17

Introduction

The partitioning of the overall molecular charge distribution into atom-centered monopole charges, while quantum mechanically ill defined, is nevertheless a technique which finds applications in several broad classes of chemical problems. Such charges can provide easily interpretable insights into the relative reactivity of particular sites within a molecule as well as provide parameters for the comparison of overall molecular reactivity. Point atomic charges are frequently used to describe the electrostatic component of the energy in force field calculations. Thus both the assignment of atomic charges in specific molecules and some assessment of the facility with which charges from smaller, known fragments can be applied to larger systems are critical to the development of transferable parameter sets for molecular mechanics and molecular dynamics.[1,2]

Since atomic charges are not formally quantum mechanical observables, they cannot be extracted from the molecular wave function in a well-defined manner, nor can they be measured experimentally. Schemes for the assignment of atomic charges therefore abound,[3-7,9,13] and methods for directly evaluating the quality of the resulting values by comparison to experiment are nonexistent. Charges can be determined empirically or by reference to a quantum mechanically derived wave function. Quantum mechanically based methods include Mulliken population analysis,[3] Bader charge partitioning,[4] natural bond orbital (NBO) population analysis,[5] and charges derived by fitting an atom-centered monopole approximation to the molecular electrostatic potential function.[6,7,9,13] The latter method has a certain appeal since the molecular electrostatic potential is a quantum mechanical observable, and therefore such charge assignment algorithms are not confined to the theoretical realm but in principle can be applied to experimental data.[7]

Just as for all schemes for atomic charge assignment, the derivation of charges from molecular electrostatic potentials is not without difficulties. In particular, it has been noted that such potential-derived charges can be conformationally dependent in ways that do not appear to reflect the changes in the molecular wave function.[1a-b,6d,8] Both the algorithm used for selecting points at which the molecular electrostatic potential will be fit and the density of points used in the fit have been suggested to influence the resultant charges.[6d,9] Recently Stouch and Williams noted that in at least one case, the least-squares data are highly correlated. The resulting numerical difficulties make it impossible to fit all the atomic charges in a molecule.[8] Obviously, both basis set[6c] and level of quantum mechanical treatment influence the outcome as well.

Solutions to several of these difficulties have been implemented. Woods has suggested using a united atom approach to moderate wild swings in hydrogen atom potential-derived charges.[10] Stouch and Williams reduced the conformational fluctuations in potential derived charges in glycerylphosphorylcholine by fixing some of the charges to "chemically reasonable" values.[8b] Breneman and Wiberg proposed that increasing the point density and careful point distribution will decrease the rotational variance observed in CHELP charges; the CHELPG version of CHELP incorporates these modifications.[6d] Urban and Famini[1d] also examine the problem of the conformational variance in these charges and suggest that CHELPG suffers from the same difficulties that it claims to avoid.

The fitting of a set of potential-derived charges $\{\vec{q}^N\}$ to molecular electrostatic potentials for an N atom molecule where the molecular electrostatic potential ($\mathbf{V}$) is known at m points is typically accomplished by a constrained least-squares procedure[11]—that is, by minimizing

$$\|\mathbf{A}^{m\times N}\vec{\mathbf{q}}^N - \vec{\mathbf{V}}^m\|$$

where $\mathbf{A}^{m\times N}$ is the least-squares matrix,[6c] $\vec{q}^N$ are the N charges in vector form, and $\vec{V}^m$ are the m values of the electrostatic potential in vector form. This procedure is subject to the constraint that the sum of the charges q_i match the actual molecular charge.[†] It has been generally, and naively, assumed that as long as the electrostatic potential is computed at more points than there are charges to be determined (i.e., $m > N$), there are sufficient data to determine all the charges by minimizing the root mean square deviation of the monopole approximation to the molecular electrostatic potential from the actual molecular electrostatic potential. If, in fact, there is a high degree of linear dependence in the least-squares matrix $\mathbf{A}$, then the

[†] Other constraints can be imposed, such as molecular symmetry or reproduction of a dipole moment, but in general only the total charge constraint is used. This enables the dipole moment to be used as a crude measure of the quality of the charges resulting from the fit. We have shown earlier (ref. 6c) that the dipole moment constraint does little to improve the rms deviation of the fits.

matrix is rank deficient and the resulting least-squares problem is ill conditioned. In case of the charges fit to the electrostatic potential, when the **A** matrix is rank deficient, there are not enough data to assign charges to all the atomic centers in the molecule. Thus, there exists an infinite number of solutions which minimize the residual norm of the molecular electrostatic potential.[‡] While the rank (r) of the least-squares matrix **A** is nominally N, an estimate ($\tilde{r}$) of the actual rank of **A** (and thus the number of charges that can be obtained) can be made by various techniques.[12]

As Stouch and Williams have noted, the seemingly overdetermined least-squares problem of fitting the electrostatic potential can in fact be highly correlated. Stouch and Williams[8b] performed a principle component analysis (PCA) on the least-squares matrix for glyceralphosphorylcholine. They find that 12 principle components describe the data in most cases, suggesting that only 12 of the 36 atomic charges in the molecular can be assigned statistically valid charges.

Once a rank estimate, and hence the number of charges which can be meaningfully assigned, has been obtained, it remains to discriminate between those atoms which can be reasonably assigned charges and those which cannot. Two general approaches can be considered: the selection of an appropriate subset and the imposition of additional constraints. Stouch and Williams[8] approach the problem by assigning fixed, "chemically reasonable" charges to various subsets of atoms (e.g., all carbon atoms) chosen on the basis of the variation in charge between conformations or by the standard deviations of the charges, and fit the remaining charges in the usual fashion. While the resulting least-squares matrix was still rank deficient even if all heavy atoms were assigned values, Stouch and Williams report that with proper selection of the subset to fit and choice of assigned charges, substantially less fluctuation in the fit charges can be achieved without an undue increase in the residual norm. The proposed methodology is not necessarily generally applicable (e.g., when a variety of conformers do not exist) and requires user input to select the charges to be fit.

[‡] Although there is an infinite set of solutions $\{\vec{x}\}$ which satisfy the least-squares equation, one can single out one of these solutions $\vec{x}_{LS}$ in which the 2-norm is at a minimum. Since computation of the 2-norm is a nontrivial task, it follows that identification of this solution is also not straightforward. See Golub and van Loan (ref. 12a), §5.5, for further information on the rank-deficient least-squares problem.

Alternatively, as Kollman et al. show, the imposition of additional constraints such as symmetry can reduce the correlation of the least-squares problem.[13] The restrained electrostatic potential (RESP) method incorporates these concepts.

The present work examines more generally the conditioning of the least-squares matrix for obtaining potential-derived charges. Our results suggest that while the least squares matrix **A** in most cases of potential-derived charges considered will be rank deficient, the ill conditioning will be more acute for some classes of molecules and types of point selection schemes than others. Such information should be of assistance in selecting molecules from which transferable potential-derived charges might be best extracted. We further examine the problem to see if extremely high point densities or larger basis sets can be used to overcome or at least improve the ill conditioning of the least-squares matrix. Finally, a well-defined, computationally efficient algorithm is presented for determining a rank estimate for the **A** matrix in potential-derived charge fitting schemes, selecting the appropriate subset of atoms to which charges can be assigned based on that rank estimate, and then refitting the selected set of charges to molecular electrostatic potential data already in hand. This scheme is well defined (i.e., it does not necessarily rely on chemical intuition for either the elimination of charges or their assignment). Further, since the value of the residual norm is generally preserved in this algorithm, the overall degradation of the fit to the molecular electrostatic potential is minimized. While this algorithm will not be applicable when it is desirable to assign a charge to each atom, since it is well defined it will be of particular use to those who wish to use the charge information as a measure of site-specific or total molecular activity. It thus provides an alternative to the imposition of additional constraints on the problem.

Methods

A modified version of CHELP[14] was used to compute the potential-derived charges. The point selection algorithm has been altered to produce as unbiased a point distribution as possible. Rather than selecting points based on a regular grid (e.g., on shells around the van der Waals radius or on a rectangular grid), we use a modification of the selection procedure of Woods et al.[9] and select

points at random within a rectangular box enclosing the molecule. This type of algorithm should thus ameliorate the unexpected asymmetries in assigned charges which can arise when a molecule's principal symmetry element is not aligned with the grid, resulting in a different distribution of points around equivalent atoms. Using a random selection of points, as opposed to a grid, can also reduce the collinearity of the matrix. Those points which fall within the van der Waals envelope of the molecule are discarded, unless otherwise noted. In this work we take the van der Waals envelope to be the surface formed by the superposition of spheres centered at the atoms. Atomic radii were taken from sphere fits to electron density surfaces.[15] The only constraint imposed on the charges is that the sum of the atomic charges be equal to the total molecular charge.

Molecular wave functions were computed using the GAUSSIAN 92[16] suite of programs and SPARTAN.[17] Structures were fully optimized at the same level as the electrostatic potential calculation, with the exception of GPC, in which the crystal structure was used.[18] The 3-21G*,[19] 6-31G*,[20] and 6-311 + G** basis sets[§] were used.

CHELP charges were obtained using the modified code described earlier for 16 molecules: formamide; ethane; acetamide; methyl acetate; dimethylphosphate; phenol; L-cysteine; *t*-butoxide; neopentane; the enol form of 2,4-pentanedione; adenine; three conformers of alanine dipeptide; a simple monosaccharide; and glycerylphosphorylcholine (GPC). These molecules were selected to represent various sizes of molecules as well as various classes of molecular shapes (e.g., disk shaped, spherical, tubular). The structures include a wide assortment of functionality, including motifs frequently encountered in biological systems, such as hydrogen bonding and amide functionalities.

We applied the singular value decomposition[12] (SVD) to the analysis of the matrices generated in the least-squares fitting procedure in CHELP. As with the principal component analyses used by Stouch and Williams, the SVD analysis can be used to give a diagnosis of the problems (e.g., the linear dependencies) in any given least-squares matrix. It has the advantage that the decomposition can also be employed in finding a new, well-conditioned least-squares matrix, and therefore a more stable solution to the problem. The SVD is a standard numerical technique and, as such, we provide only a brief introduction to it in this article. We refer the interested reader to the literature cited for further information.

The SVD is an orthogonal decomposition of an $M \times N$ matrix that satisfies the following:

$$\mathbf{A} = \mathbf{USV}^{\mathrm{T}}$$

where **A** is the least-squares matrix,[6c] and **U** and **V** are orthogonal matrices—that is, they satisfy

$$\sum_{i=1}^{M} U_{ik}U_{in} = \delta_{kn} \quad \text{for} \quad 1 \le k \le N; 1 \le n \le N$$

$$\sum_{j=1}^{N} V_{jk}V_{jn} = \delta_{kn} \quad \text{for} \quad 1 \le k \le N; 1 \le n \le N$$

The diagonal entries of **S**, $\{s_1, s_2, \ldots, s_N\}$, are the singular values of the least-squares matrix **A**. For N charges there will be N singular values. The condition number of **A** is defined by the ratio of the largest ($s_{\max}$) to the smallest singular value ($s_{\min}$). When the matrix **A** is singular, the condition number is infinite. If the condition number is very large, the matrix is ill conditioned and thus rank deficient. An estimate of the rank of **A** can be found by examining the ratios $s_i/s_{\max}$ for all N values. For each i where the ratio is less than the precision of the data, the rank is reduced by one. It can easily be seen that as we eliminate the smaller singular values, the condition number for the matrix increases. When a sufficient number of singular values has been eliminated, we have a new, better conditioned and smaller matrix **A** to work with. In this work we have reduced the rank of the matrix by one for each singular value where $s_i/s_{\max}$ was less than 0.0001. This assumes that the error in the calculated electrostatic potential is on the order of 10^{-7} hartrees.

As we noted earlier, the SVD can be used not only to find the maximum size of the properly conditioned least-squares problem (i.e., the number of charges), but it can also be used to find the solution to the smaller, better conditioned problem. To do this, it is necessary to decide which parameters can be fit and which must be discarded; $\tilde{r}$ or fewer parameters may be retained. Several algorithms based on the SVD exist for selecting a subset of parameters which can be fit acceptably from the extant data. We use the one developed by Golub, Klema, and Stewart.[‖] This algorithm is designed to preserve, as much as

[§] See ref. 21. The 6-311 + G** basis includes both diffuse and polarization functions on heavy atoms and *p*-type polarization functions on hydrogen.

[‖] See ref. 12a, p. 571.

possible, the residual norm (i.e., the root mean square deviation of the monopole potential from the true electrostatic potential). We feel that this scheme is more likely than some of the others to give good dipole moments and good reproduction of the molecular electrostatic potential, *vide infra*.

The new subset is selected by first computing the SVD of our least-squares matrix **A** to obtain the orthogonal matrix **V** and $\tilde{r}$. The QR (ref. 12) scheme with column pivoting is then applied to $\mathbf{V}^T$

$$\mathbf{Q}^T\mathbf{V}^T\mathbf{P} = \begin{pmatrix} \mathbf{R}_1 & 0 \\ 0 & \mathbf{R}_2 \end{pmatrix}$$

where $\mathbf{R}_1 \in \Re^{\tilde{r}x\tilde{r}}$ and $\mathbf{R}_2 \in \Re^{(N-\tilde{r})x(N-\tilde{r})}$

to generate a diagonal permutation matrix **P** which can be used to reorder the data in **A** such that the first $\tilde{r}$ parameters in **A** correspond to the "best" subset, which will be refit to the data in **A**.

$$\mathbf{AP} = \begin{pmatrix} \mathbf{B}_1 & 0 \\ 0 & \mathbf{B}_2 \end{pmatrix}$$

where $\mathbf{B}_1 \in \Re^{mx\tilde{r}}$ and $\mathbf{B}_2 \in \Re^{mx(N-\tilde{r})}$

$\mathbf{B}_1$ now contains the data needed to fit a new set of charges which will still satisfy the original least-squares conditions but will not include charges designated as unassignable by the SVD analysis. The new set of charges $\{q\}$ can be obtained by minimizing

$$\|\mathbf{B}_1^{mx\tilde{r}}\vec{\mathbf{q}}^{\tilde{r}} - \vec{\mathbf{V}}^m\|$$

subject to the constraints of the original problem.

CHELP was modified to incorporate the foregoing analysis and refitting scheme; the modified version is referred to as CHELP-SVD later.[¶] The LINPACK routines were used to obtain the SVD and QR.[**]

Results and Discussion

Rank estimates arrived at using the SVD are summarized in Table I for the molecules formamide; ethane; acetamide; methyl acetate; dimethylphosphate; phenol; L-cysteine; *t*-butoxide; neopentane; the enol form of 2,4-pentanedione; adenine; three conformers of alanine dipeptide; glucose; and glycerylphosphorylcholine (GPC). Molecules and their numbering schemes are shown in Figure 1. The CHELP charges were fit to HF/3-21G* molecular electrostatic potentials. The total number of points at which the electrostatic potential was computed varied between 10^2 and 10^4. When a total of 1000 points was used, the point density was O(10) points/Å^3 (e.g., 37 points/Å^3 for formamide, 20 points/Å^3 for adenine). This is roughly the point density that is standard in CHELPG.[6d] Points within the van der Waals radius were excluded, as were points outside an envelope three times the van der Waals radius.

The data in Table I clearly show that rank deficiency is not limited to very large molecules, such as GPC, although the problem generally increases

TABLE I.
SVD Rank Estimates for CHELP Least-Squares Fit of Charges.[a]

Molecule	Full Rank[b]	SVD Rank Estimate[c]
Formamide	7	6
Ethane	9	8
Acetamide	10	8
Methyl acetate	12	10
Dimethylphosphate	14	11
Phenol	14	11
L-cysteine	15	12
t-Butoxide	15	12
2,4-Pentanedione (enol form)	16	13
Adenine	16	12
Neopentane	18	14
Alanine dipeptide (7a)	23	17
Alanine dipeptide (7e)	23	18
Alanine dipeptide (tr)	23	18
Glucose	25	18
GPC	37	25

[a] Fits are to HF/3-21G(*) electrostatic potential calculated at 1000 points in a shell between 1.0 and 3.0 times the van der Waals surface of the molecule.
[b] Full rank of least-squares matrix is the number of charges to be fit (in these cases, the number of atoms in the molecule) plus the number of restraints imposed. In the tabulated calculations, the only constraint is that the total molecular charge equals the sum of the atomic charges.
[c] Rank estimate taken as number of singular values for which the ratio of s_i to s_{max} does not exceed 10^{-4}. See text for additional details.

[¶] CHELP-SVD is available on the Internet via gopher from gopher.brynmawr.edu. Contact mfrancl@cc.brynmawr.edu for details.

[**] Modifications of routines dqrdc.f and dqrsl.f were used. Further information and routines can be obtained by sending the command "send index" to netlib@research.att.com on the Internet.

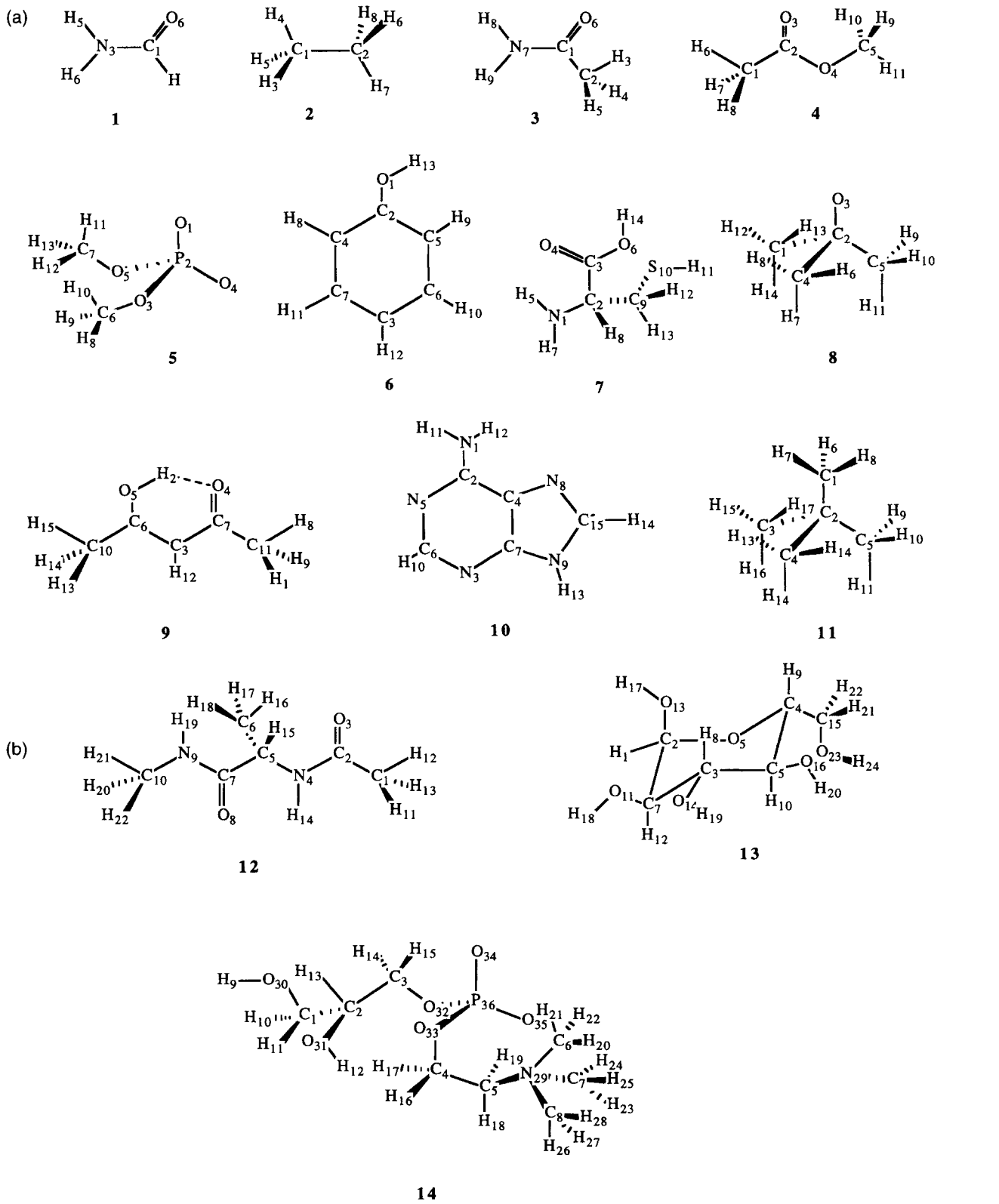

FIGURE 1. Numbering schemes for **1**, formamide; **2**, ethane; **3**, acetamide; **4**, methyl acetate; **5**, dimethylphosphate; **6**, phenol; **7**, L-cysteine; **8**, *t*-butoxide; **9**, 2,4-pentanedione (enol form); **10**, adenine; **11**, neopentane; **12**, alanine dipeptide; **13**, glucose; **14**, GPC.

with increasing molecular size. For example, the SVD reveals that the least-squares matrix for ethane is not quite full rank (eight); only seven of the eight atoms can thus be meaningfully assigned charges in this molecule. (Note that the number of atoms which can be fit is one *less* than the estimated rank, since the constraint that the total charge equal the molecular charge cannot be released.) We have examined formamide, for comparison with the work of Breneman and Wiberg.[6d] We find that even here the least-squares matrix is not full rank and that one or more of the charges is undeterminable by these methods. Acetamide behaves similarly; with an SVD rank estimate of 8, all but two of the charges can be assigned. The least-squares matrix for phenol is estimated by SVD to have a rank of 11, as compared to the full rank of 14, suggesting that just over 75% of the charges can be meaningfully specified in phenol. The SVD results for neopentane suggest similarly that about 75% of charges can be fit to the data at hand. Larger, less symmetrical molecules are not expected to fare as well. Only about 70% of the charges can be fit to the extant data in the 7e conformer of alanine dipeptide. Our results for GPC are consistent with those of Stouch and Williams; we find that 60% of the charges can be assigned.

As Figure 2 shows, the percentage of charges which can be fit decreases with increasing molecular size, albeit not smoothly. This decline is not unexpected when one considers classical electrostatics. Classically, all the charge on an object is found to be on the surface of the object. The electrostatic potential calculated outside the molecular van der Waals surface will tend to reflect this proclivity. For example, Stouch and Williams have noted that atoms buried in the interior of GPC seem to be less important to the fit than exterior atoms.[8b] As a larger percentage of the specified charge sites (i.e., atoms) is found away from that surface, a larger percentage of the charges will become unfitable. To a first approximation, the percentage of assignable charges should fall off as $1/\sqrt[3]{r}$.[††] This curve is shown superimposed on the data in Figure 2. We suggest that fits to electrostatic potentials are a less reliable way of deriving chemically meaningful charges for very large molecules.

[††] The $1/\sqrt[3]{r}$ behaviors assume that the molecules are spherical and that the total number of atoms is proportional to the volume of the sphere, while the number of exterior charges is proportional to the surface area of the molecular sphere.

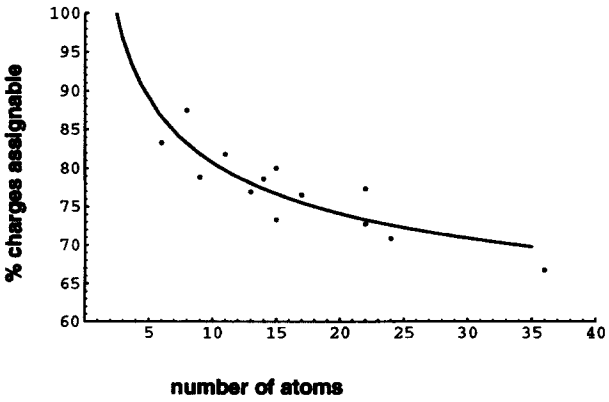

FIGURE 2. Percentage of charges that can be assigned as a function of the total number of atoms in the molecule. Data shown are from SVDs of CHELP matrices at the HF/3-21G(*) level using 1000 points.

An evaluation of rank deficiency and an algorithm to select the meaningful charge data are useful because they prevent both the transfer of meaningless charges to larger systems and the analysis of chemical phenomena based on what are essentially random numbers. However, there are certainly cases in which the assignment of charges to all of the atoms is desirable. What can be done to ameliorate the rank deficiency? Breneman and Wiberg[6d] have suggested that higher point densities are helpful. The data show, in direct contrast to this popular notion, that increasing the point density will generally not be effective in increasing the number of charges which can be meaningfully assigned in a given molecule. Some of the difficulties are inherent in the problem; they are not artifacts of either the choice of, nor the number of, points selected.

The effect of increasing point density on the number of assignable charges is clearly shown in Table II. After a certain point (roughly 10^2 points), doing so has no effect on the rank, even for small molecules such as formamide. Since the cost of least-squares algorithms such as CHELP is proportional to the number of points at which the molecular electrostatic potential is calculated, there is a substantial time penalty to pay for computing more points than needed.

The foregoing electrostatic argument also implies that tightening the surface, bringing the specified charge sites closer to the exclusion surface, could be effective in increasing the rank of the least-squares matrix. This in fact appears to be the case, as the data in Figure 3 suggest. Drawing the surface of the molecule at 0.5 times the van der

TABLE II.
SVD Rank Estimates for CHELP Least-Squares Fit[a] as a Function of Total Number of Points.

Molecule	100	1000	10,000
Formamide	6	6	5
Ethane	7	8	7
Acetamide	8	8	7
Methyl acetate	10	10	9
Dimethylphosphate	11	11	10
Phenol	11	11	10
L-cysteine	11	12	11
t-butoxide	11	12	11
2,4-Pentanedione (enol form)	13	13	12
Adenine	11	12	11
Neopentane	13	13	13
Alanine dipeptide (7a)	17	17	17
Alanine dipeptide (7e)	16	18	17
Alanine dipeptide (tr)	20	18	17
Glucose	17	18	17
GPC	22	25	24

[a]Fits are to HF/3-21G(*) electrostatic potential data calculated at points selected to be within a shell between 1.0 and 3.0 times the van der Waals surface of the molecule.

Waals radius in general increases the SVD rank estimate, enabling additional charges to be assigned. The principal drawback to using this surface is the poor reproduction of the electrostatic potential that it can yield, which is due to both the breakdown of the monopole approximation and the failure of the perturbation approximation used to calculate the molecular electrostatic potential. Root mean square (rms) deviations of the monopole potential from the true potential can be extremely high when a surface this tight is used (on the order of 25 kcal/mol). Loosening the surface slightly, 0.75 times the van der Waals radii, yields a better balance between high rank and low rms deviation. The rms deviations are on the order of 5 kcal/mol when the 0.75 surface is used, in comparison with the range of 0.5 to 2.0 kcal/mol typical of the fits from the 1.0 exclusion surface. At this point, the perturbation methods are adequate to the task,[21] although the monopole approximation still falters in this region.

Table III shows the atoms for each molecule which the Golub algorithm has selected as being most likely to have meaningful CHELP charges. Based on the electrostatic argument presented earlier, we expected to find that charges on interior atoms are less meaningful than those of atoms on the accessible surface of a molecule. This turns out not to be the case. For the molecules considered here, the choices appear not to be critical. The QR algorithm merely selects the last $(N - \tilde{r})$ charges to exclude. This suggests that if one is interested in reproducing the electrostatic potential accurately using the monopole approximation, one should use the QR selected subset. If one desires charges for use in the analysis of a chemical problem, one can select up to a maximum of $\tilde{r}$ charges to fit without substantial penalty, in much the fashion that Stouch and Williams suggested.

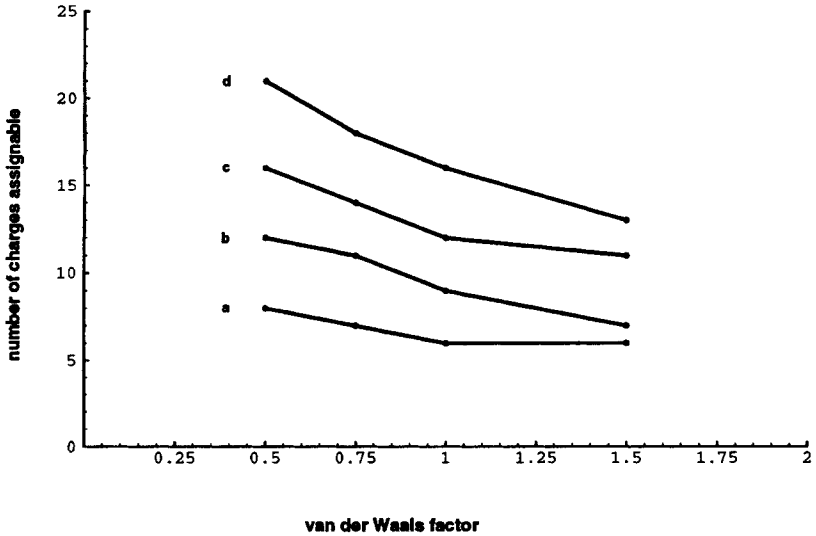

FIGURE 3. Number of charges assignable as a function of the size of the envelope surrounding the molecule. Data shown are from SVDs of CHELP matrices at the HF/3-21G(*) level.

TABLE III.
Best Set of Charges to Fit[a] Using the CHELP-SVD Algorithm.

Molecule	Number of Charges to Fit[b]	QR Selected Subset[c]
Formamide	5	{C1,O2,N3,H4,H5}
Ethane	7	{C1,C2,H3,H4,H5,H6,H7}
Acetamide	7	{C1,C2,H3,H4,H5,O6,N7}
Methyl acetate	9	{C1,C2,O3,O4,C5,H6,H7,H8,H9}
Dimethylphosphate	10	{O1,P2,O3,O4,O5,C6,C7,H8,H9,H19}
Phenol	10	{O1,C2,C3,C4,C5,C6,C7,H8,H9,H10}
L-cysteine	11	{N1,C2,3,O4,H5,O6,H7,H8,C9,S10,H11}
t-Butoxide	11	{C1,C2,O3,C4,C5,H6,H7,H8,H9,H10,H11}
2,4-Pentanedione (enol form)	12	{H1,H2,C3,O4,O5,C6,C7,H8,H9,C10,C11,H12}
Adenine	11	{N1,C2,N3,C4,N5,C6,C7,N8,N9,H10,H11}
Neopentane	13	{C1,C2,C3,C4,C5,H6,H7,H8,H9,H10,H11,H12,H13}
Alanine dipeptide (7a)	16	{C1,C2,O3,N4,C5,C6,C7,O8,N9,C10,H11,H12,H13,H14,H15,H16}
Alanine dipeptide (7e)	17	{C1,C2,O3,N4,C5,C6,C7,O8,N9,C10,H11,H12,H13,H14,H15,H16,H17}
Alanine dipeptide (tr)	17	{C1,C2,O3,N4,C5,C6,C7,O8,N9,C10,H11,H12,H13,H14,H15,H16,H17}
Monosaccharide	17	{H1,C2,C3,C4,C5,O6,C7,H8,H9,H10,O11,H12,O13,O14,C15,O16,H17}
GPC	24	{C1,C2,C3,C4,C5,C6,C7,C8,H9,H10,H11,H12,H13,H14,H15,H16,H17, H18,H19,H20,H21,H22,H23,H24}

[a]Fits are to HF/3-21G(*) electrostatic potential calculated at 1000 points in a shell between 1.0 and 3.0 times the van der Waals surface of the molecule.
[b]Rank −1. Rank estimate from SVD.
[c]See Figure 1 for numbering scheme.

Charges from the full CHELP analysis are compared to charges from the subset selected by CHELP-SVD, as well as to a subset selected using "chemical intuition" (referred to as the CHELP-SVD/user charges) in Table IV. It is expected that charges assigned to any given atom by the three methods will differ. Charges on exterior atoms should be largely unaffected; charges on atoms that are farther from the molecular surface may show substantial differences.

It can be seen that in most cases, using the CHELP-SVD charges to model the molecular electrostatic potential does not substantially affect the quality of the fit. In general, the CHELP-SVD/user molecular electrostatic potentials are closer to those from CHELP-SVD. The rms change between the CHELP fits and the CHELP-SVD fits is 4.3 kcal/mol, while the rms for CHELP compared to the CHELP-SVD/user fits is only 1.6 kcal/mol. Charges from CHELP-SVD are generally less illuminating than those from the CHELP-SVD/user model, since the choice of atoms to fit is essentially random (merely a function of the order in which the atoms were input to either Spartan or GAUSSIAN 92).

Comparing the CHELP-SVD/user-selected charges on acetamide with those from standard CHELP shows that the buried charge (in this case C2) is the only one substantially affected. CHELP-SVD/user fits a charge of 0.02 to this carbon, more in line with general chemical notions than the −0.69 that full CHELP fit predicts. A similar trend is seen in methyl acetate, where the charges on the two hidden carbons (C1 and C5) make more chemical sense. CHELP-SVD/user predicts C1 to be nearly neutral (0.02) rather than very negative (−0.90); while C5, which one might expect to be $\delta+$, has a CHELP-SVD/user charge of 0.38 compared to more than twice that predicted by CHELP. Comparable improvements in the prediction of the charges in buried atoms can be seen in glucose. CHELP predicts the charges on C2 and C4 to be 0.36 and 0.09, respectively. While the trend is certainly what one might expect from simple electronegativity arguments (C2 more positive than C4), the nearly neutral character of C4 is unanticipated. CHELP-SVD/user charges do not affect the trend (0.64 compared to 0.28), but the $\delta+$ nature of C2 is now revealed. The CHELP-SVD/user charges on the carbons of neopentane are also

an improvement on the CHELP model; the bonds are predicted to be much less polar (Δq for CHELP-SVD/user is only 0.43 as compared to 1.15 with CHELP) and the methyl carbons are nearly neutral.

One of the recognized failures of CHELP has been that it can assign different charges to atoms that are equivalent by symmetry. It has been suggested[6d] that this is due to a low density of points and/or the asymmetry of the points selected for the fit. We show here that it is not necessarily due to either factor, but can be a result of attempts to fit more charges than one has adequate data to assign. The CHELP charges on the equivalent carbon atoms (C6 and C7) in dimethylphosphate differ by almost 0.2 e^-; the CHELP charges on the two oxygens equivalent by symmetry (O3 and O5) also differ significantly (-0.75 compared to -0.70). On the other hand, CHELP gives nearly equal charges to the two nonequivalent oxygens (O1 and O4). CHELP-SVD/user yields identical or nearly identical ($\Delta q \leq .01$ e^-) charges on both the C6/C7 and O3/O5 pairs while assigning distinctly different charges to the nonequivalent set of O1/O4. A

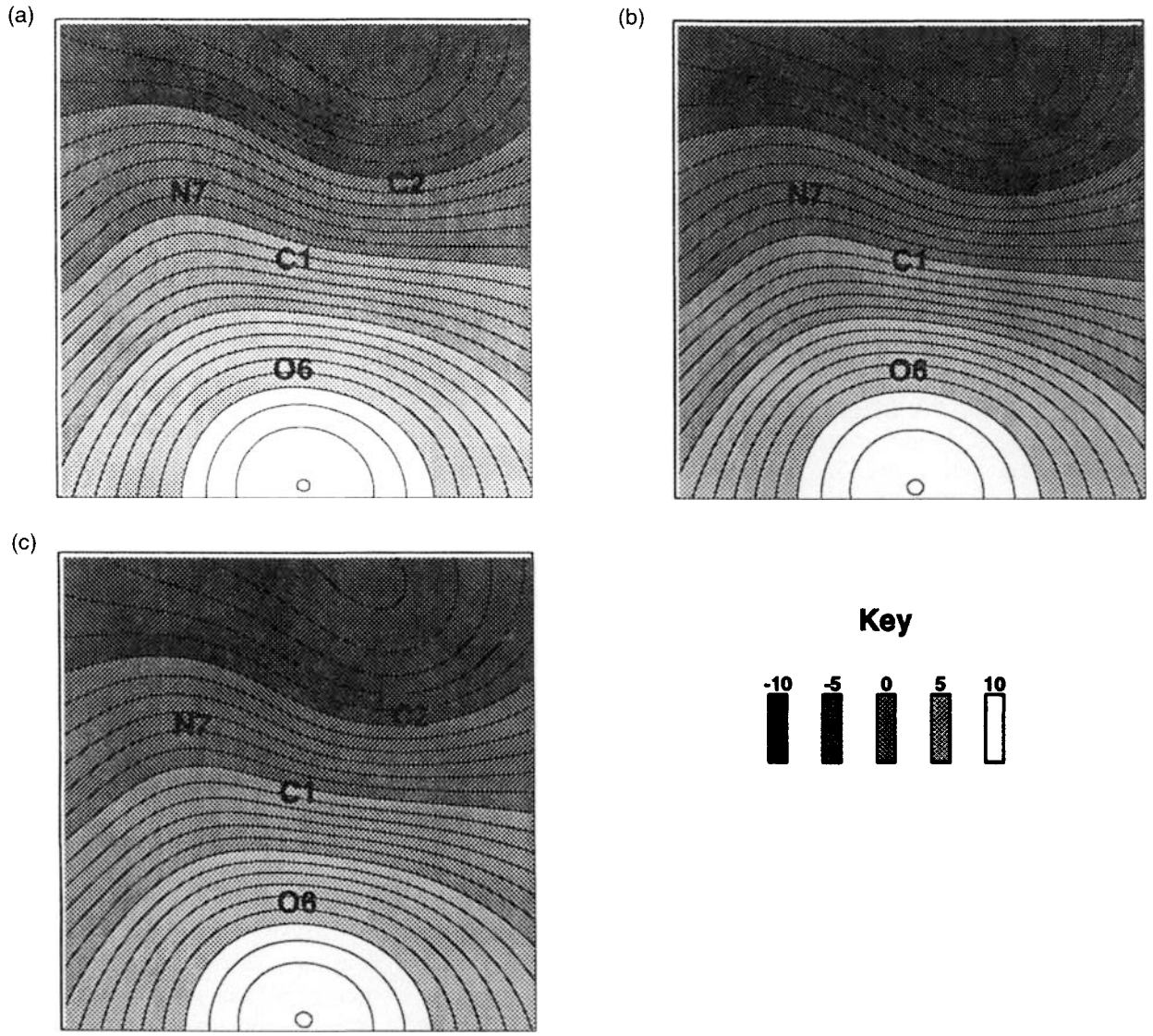

FIGURE 4. Contour maps of the molecular electrostatic potential for acetamide at the HF/3-21G(*) level calculated (a) from the full molecular wave function; (b) using the monopole approximation and the CHELP charges; and (c) using the monopole approximation and the CHELP-SVD/user charges. Shading indicates approximate value of the potential in the region.

similar improvement in the symmetry of the charges is seen with both *t*-butoxide and neopentane. The CHELP charges on the three carbon atoms of *t*-butoxide which are equivalent by symmetry (C1, C4, and C5) differ by roughly 0.05 e^-, while the CHELP-SVD/user charges differ by less than 0.01 e^-. A more dramatic improvement is seen in neopentane, in which the charges on the four equivalent carbon atoms (C1, C3, C5, and C5) differ by as much as 0.08 e^- using the CHELP model, while the CHELP-SVD/user model predicts differences less than 0.01 e^-.

When selecting a subset of charges to be fit, one must take care not to reduce the subset too far. For example, if one selects a subset of eight atoms (O1, C2, C3, C4, C5, C6, C7, and H13) as an appropriate set for phenol, the resulting charges are not what is intuitively expected for this system. C2, for example, is not predicted to be $\delta+$, but instead bears a significant negative charge. The rms deviation points up the poor quality of the fit (it is nearly 5.00 kcal mol^{-1}), as does the qualitative comparison between the true molecular electrostatic potential and the CHELP-SVD/user potential. Charges from fits with high rms deviations ($\geq$ 5 kcal mol^{-1}), no matter what the model, should be suspect. Since the SVD suggests that up to 10 charges can be fit, adding two more atoms to the subset should and does alleviate the problem, as the data in Table IV show.

Figures 4 and 5 illustrate the differences in the predicted electrostatic potentials between the full CHELP and CHELP-SVD/user models. HF/3-21G molecular electrostatic potentials for acetamide are compared with those from CHELP and CHELP-SVD/user. As one might expect from the rms deviations, both CHELP and CHELP-SVD/user do a good job of reproducing the qualitative features

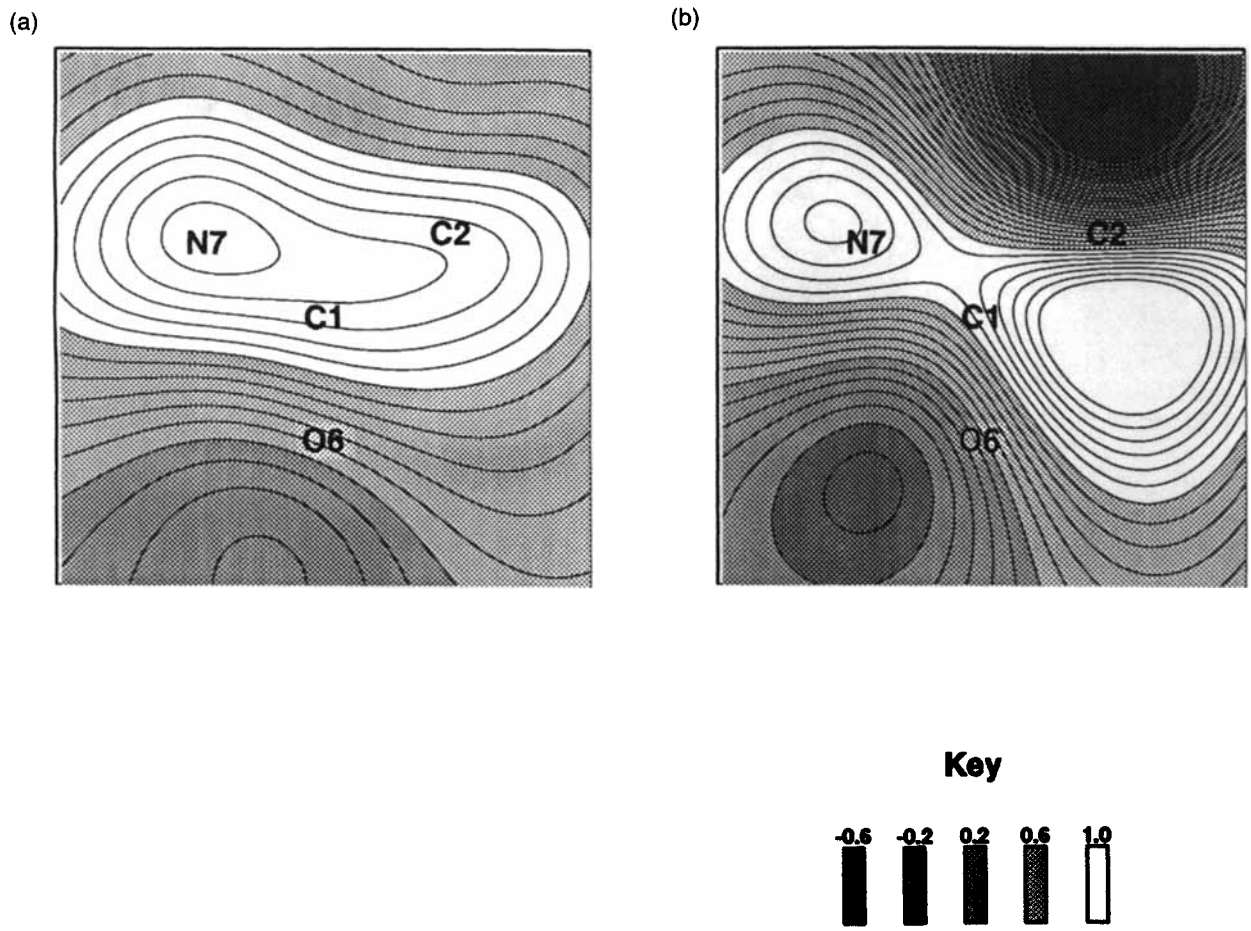

FIGURE 5. Contour maps of the difference between the molecular electrostatic potential from the full wave function and that from (a) the monopole approximation using the CHELP charges and (b) the monopole approximation using the CHELP-SVD/user charges for acetamide at the HF/3-21G(*) level. Shading indicates approximate value of the potential in the region.

TABLE IV.
Comparison of Charges for Selected Atoms from CHELP Fits to Full Set of Atoms and from Fits to Subsets from CHELP-SVD.

Molecule	Atom[a]	CHELP Charges[b]	CHELP-SVD Charges[c]	User-Selected Subset	Charges for User's Subset[d]
Acetamide	C1	1.02	0.10	{C1,C2,O6,N7,H8,H9}	0.79
	C2	−0.69	−0.05		0.02
	O6	−0.66	−0.45		−0.63
	N7	−1.12	0.25		−1.00
	H8	0.45	—		0.43
	H9	0.47	—		0.39
	rms[e]	0.86	3.84		1.26
Methyl acetate	C1	−0.90	−0.77	{C1,C2,O3,O4,C5}	0.02
	C2	1.17	1.10		0.89
	O3	−0.68	−0.66		−0.64
	O4	−0.61	−0.64		−0.65
	C5	0.16	0.35		0.38
	rms[e]	1.26	1.11		1.88
Dimethylphosphate	P2	1.72	1.75	{O1,P2,O3,O4,O5,C6,C7}	1.82
	O1	−0.92	−0.91		−0.91
	O4	−0.93	−0.94		−0.97
	O3	−0.75	−0.78		−0.70
	O5	−0.70	−0.68		−0.70
	C6	0.54	0.68		0.23
	C7	0.36	0.23		0.24
	rms[e]	1.61	1.78		1.88
t-Butoxide	O3	−0.44	−1.05	{C1,C2,O3,C4,C5}	−1.00
	C2	1.10	1.09		1.00
	C1	−0.44	−0.31		−0.34
	C4	−0.48	−0.51		−0.33
	C5	−0.49	−0.56		−0.32
	rms[e]	1.00	1.07		1.71
2,4-Pentanedione (enol)				{H2,C3,O4,O5,C6, C7,C10,C11}	
	C3	−1.20	−1.29		−1.18
	C6	0.97	0.82		0.74
	C7	1.09	1.27		0.99
	C10	−0.67	0.06		0.08
	C11	−0.64	−0.89		−0.02
	O5	−0.76	−0.78		−0.76
	O4	−0.71	−0.76		−0.74
	H2	0.54	0.56		0.57
	rms[e]	0.90	1.49		1.50
Adenine				{N1,C2,N3,C4,N5, C6,C7,N8,N9,C15}	
	N1	−0.86	0.08		0.26
	C2	−0.83	0.45		0.26
	N5	−0.84	−0.90		−0.72
	C6	0.66	0.99		0.72
	N3	−0.83	−0.76		−0.63
	C7	0.72	0.05		0.17
	N9	−0.53	0.53		0.21
	C15	0.24	—		0.31
	N8	−0.50	−0.27		−0.48
	C4	−0.21	−0.22		−0.10
	rms[e]	1.49	4.97		5.00

TABLE IV.
(continued)

Molecule	Atom[a]	CHELP Charges[b]	CHELP-SVD Charges[c]	User-Selected Subset	Charges for User's Subset[d]
L-cysteine	N1	−1.19	−0.98	{N1,C2,C3,O4,H5, O6,H7,H11,H14}	−1.13
	H5	0.47	0.29		0.46
	H7	0.46	0.52		0.45
	C2	0.53	−0.51		0.36
	C3	0.77	0.65		0.81
	C9	−0.14	0.32		0.15
	O6	−0.80	0.05		−0.80
	H14	0.54	—		0.54
	O4	−0.62	−0.44		−0.62
	S10	−0.39	−0.36		−0.46
	H11	0.22	0.18		0.23
	rms[e]	2.00	4.61		2.14
Phenol				{O1,C2,C3,C4,C5, C6,C7,H8,H9,H13}	
	C2	0.64	0.57		0.95
	C3	−0.43	0.15		−0.16
	C4	−0.50	−0.67		−0.96
	C5	−0.51	−0.20		−1.00
	C6	0.08	−0.62		0.37
	C7	0.08	0.22		0.38
	O1	−0.70	−0.22		−0.73
	H13	0.43	—		0.44
	rms[e]	1.10	6.91		2.80
Neopentane	C1	−0.38	−0.49	{C1,C2,C3,C4,C5}	−0.09
	C2	0.69	0.64		0.34
	C3	−0.38	−0.11		−0.09
	C4	−0.41	−0.16		−0.08
	C5	−0.46	−0.62		−0.08
	rms[e]	0.94	6.91		1.56
Alanine dipeptide (7a)				{C1,C2,O3,N4,C5,C7,O8, N9,C10,H14,H19}	
	C1	−0.40	−0.47		0.00
	C2	0.97	0.60		0.81
	O3	−0.66	−0.38		−0.61
	N4	−0.86	−0.87		−0.77
	H14	0.33	0.40		0.32
	C5	0.37	0.69		0.26
	C7	0.85	0.23		0.82
	O8	−0.68	−0.55		−0.67
	N9	−0.86	−0.07		−0.69
	H19	0.32	—		0.30
	C10	0.59	0.14		0.24
	rms[e]	2.69	2.43		2.59

TABLE IV.
(continued)

Molecule	Atom[a]	CHELP Charges[b]	CHELP-SVD Charges[c]	User-Selected Subset	Charges for User's Subset[d]
Glucose				{C2,C3,C4,C5,O6,C7, O11,O12,H13,O14, O16,H17,H18,H19, H20,O23,H24}	
	C2	0.36	0.89		0.64
	C3	0.36	1.33		0.35
	C4	0.09	1.11		0.28
	C5	0.29	−1.42		0.36
	C7	0.26	−0.33		0.28
	O6	−0.58	−0.69		−0.66
	O11	−0.73	−0.11		−0.74
	H18	0.44	—		0.43
	O13	−0.78	−0.44		−0.82
	H17	0.51	—		0.50
	O14	−0.81	−0.23		−0.80
	H19	0.48	—		0.48
	O16	−0.81	0.12		−0.85
	H20	0.50	—		0.51
	O23	−0.79	—		−0.71
	H24	0.48	—		0.44
	rms[e]	1.10	6.90		1.46
GPC				{C1,C2,C3,C4,C5,C6,C7, C8,H9,H12,N29,O30, O31,O32,O33,O34, O35,P36}	
	C1	0.25	0.00		0.27
	C2	0.52	0.84		0.27
	C3	0.20	−2.83		0.35
	P36	1.46	—		1.71
	O32	−0.53	—		−0.64
	O33	−0.68	—		−0.68
	O34	−0.82	—		−0.88
	O35	−0.87	—		−0.89
	C6	−0.45	1.48		0.23
	N29	0.35	—		0.07
	C5	−0.29	0.14		0.16
	C4	0.41	−1.15		0.30
	C7	−0.47	0.80		0.20
	C8	−0.14	−0.01		0.13
	O30	−0.79	—		−0.75
	H9	0.48	0.00		0.45
	O31	−0.88	—		−0.71
	H12	0.47	−0.02		0.40
	rms[e]	1.08	10.1		2.18

[a] See Figure 1 for numbering scheme.
[b] Fit to full set of atoms, constrained to reproduce molecular charge. All fits are to HF / 3-21G(*) data; 1000 points selected between 1.0 and 3.0 times the van der Waals surface of the molecule.
[c] Fit to subset selected by CHELP-SVD. See Table III for results.
[d] Charges fit to subset of atoms selected by user given in previous column. See text for additional details.
[e] Root mean square deviation of fit electrostatic potential from actual HF / 3-21G(*) potential in kcal mol^{-1}.

of the molecular electrostatic potential (Fig. 4). The difference plots in Figure 5 show that the molecular electrostatic potential in the area between C1 and C2 is somewhat better predicted by the CHELP-SVD/user charges than by the CHELP charges, while the area behind C2 is better predicted by CHELP. At all points in this plane, neither the CHELP-SVD/user nor the CHELP model molecular electrostatic potentials differ by more than 1 kcal/mol from the full HF/3-21G results.

Figures 6 and 7 show the extremely poor reproduction of the molecular electrostatic potential when the subset is overly restricted; in this case eight atoms have been used for phenol instead of ten. The CHELP-SVD/user molecular electrostatic potential (Fig. 6c) is qualitatively different from the HF/3-12G potential. The CHELP-SVD/user

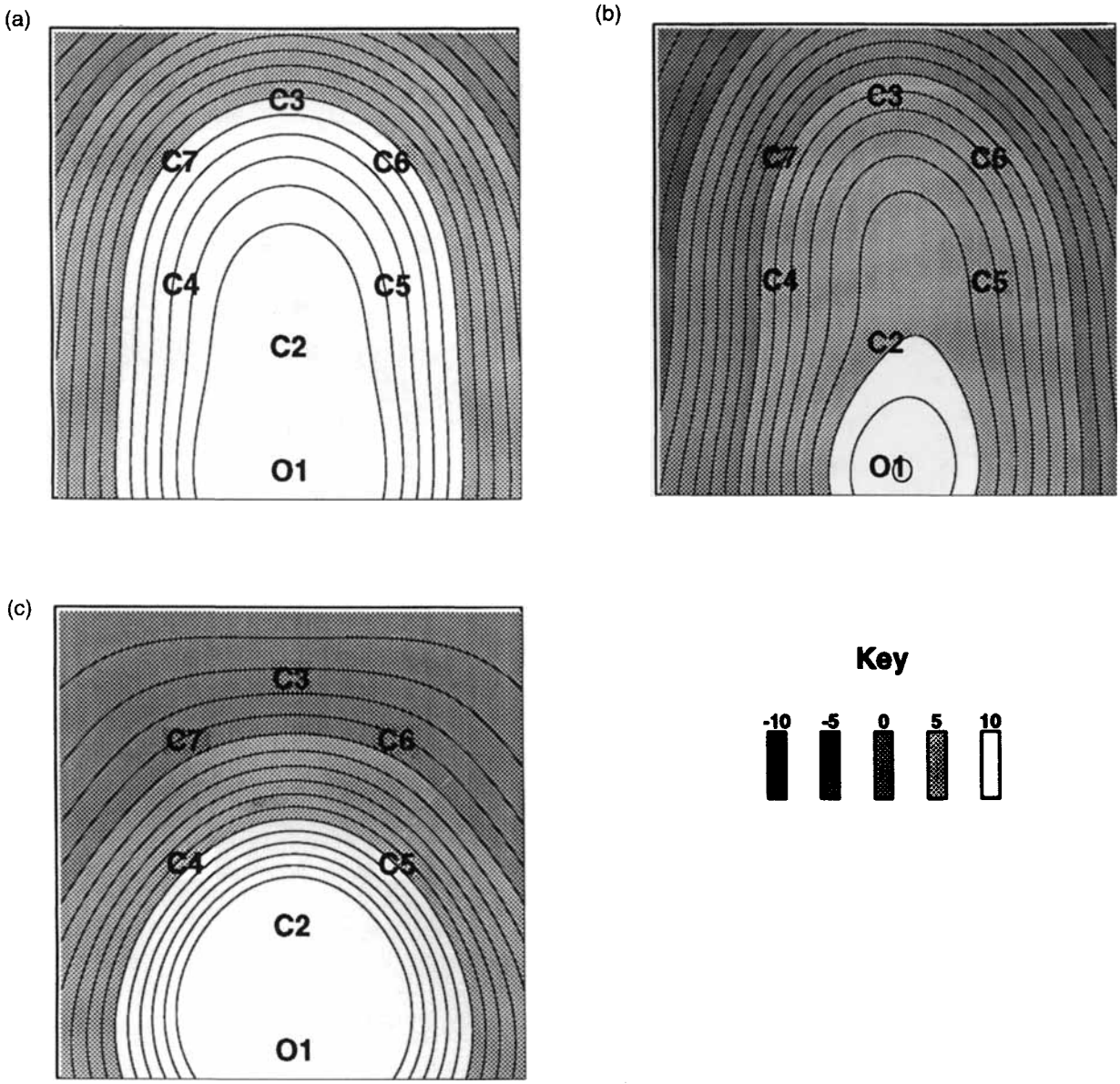

FIGURE 6. Contour maps of the molecular electrostatic potential for phenol at the HF/3-21G(*) level calculated (a) from the full molecular wave function; (b) using the monopole approximation and the CHELP charges; and (c) using the monopole approximation and an overly constrained set of CHELP-SVD/user charges. Shading indicates approximate value of the potential in the region.

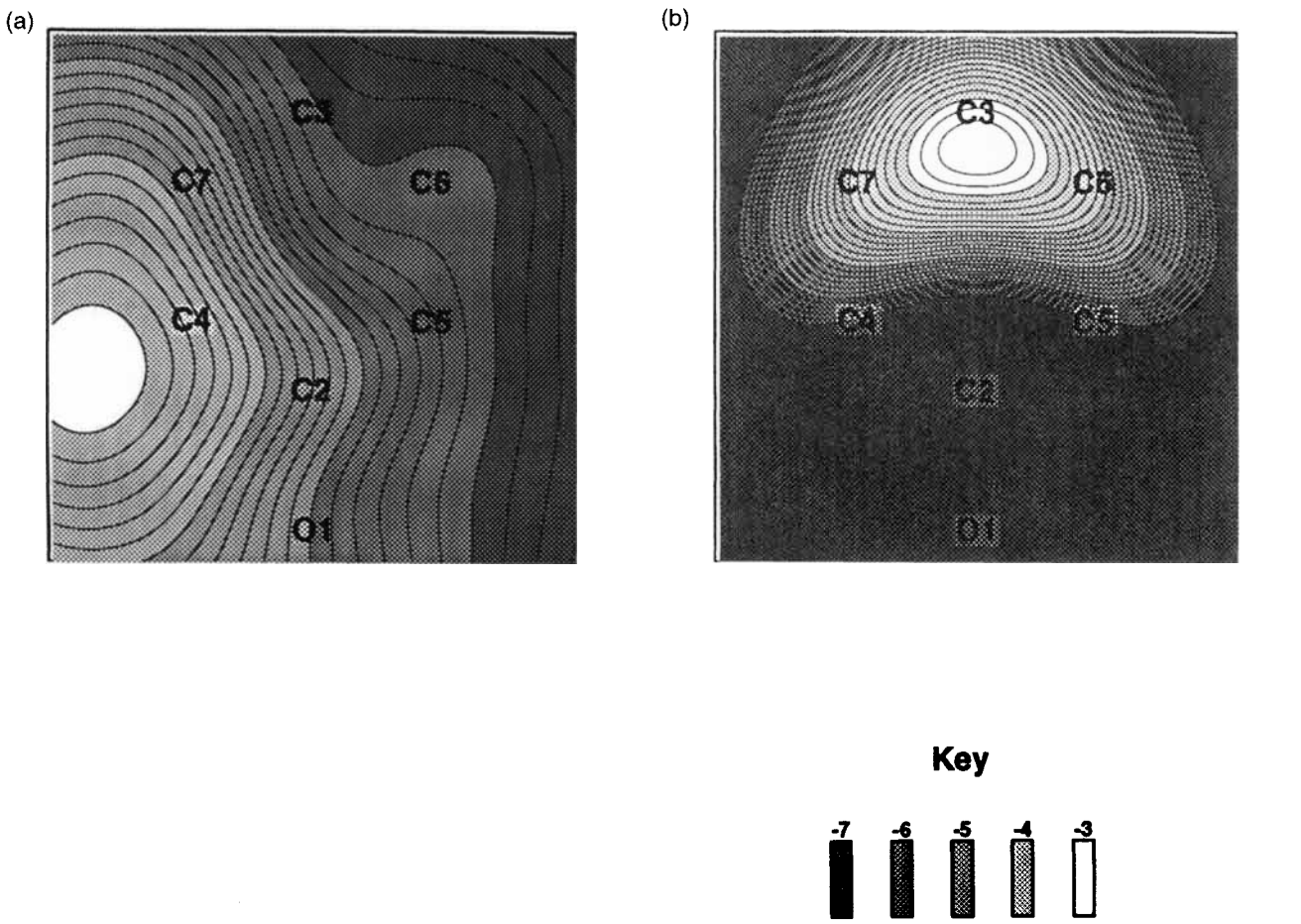

FIGURE 7. Contour maps of the difference between the molecular electrostatic potential from the full wave function and that from (a) the monopole approximation using the CHELP charges and (b) the monopole approximation using an overly constrained set of CHELP-SVD / user charges for phenol at the HF / 3-21G(*) level. Shading indicates approximate value of the potential in the region.

molecular electrostatic potential within the ring differs by as much as 5 kcal/mol from the true potential, while differences around the phenolic oxygen are even larger than that. As we noted earlier, fits with large rms are not reliable, either as quantitative or qualitative descriptions of the molecular electrostatic potential.

Stouch and Williams suggest that larger basis sets might enhance the assignment of accurate charges. As we noted earlier, the use of *accurate* in this context is misleading. The data in Table V suggest that increasing the basis set from 3-21G(*) to 6-311 + G* does not substantially change the rank of the least-square matrices. Of course, the change in basis set does affect the magnitude of the charges (a matter addressed in an earlier article[6c]) and may therefore produce a better set for some purposes.

TABLE V.
SVD Rank Estimates[a] for CHELP Least-Squares Fit of Charges Fit to 1000 Points Using Data from HF / 3-21G* and HF / 6-311 + GCalculations.**

Molecule	3-21G(*)	6-311 + G**
Acetamide	8	8
Neopentane	14	14
Alanine dipeptide (tr)	18	18
Alanine dipeptide (7a)	17	17

[a]Rank estimate taken as number of singular values for which the ratio s_i to s_{max} does not exceed 10^{-4}. See text for additional details.

Conclusions

Singular value decomposition of the linear least-squares matrices used in fitting atom-based

monopoles to molecular electrostatic potentials provides a tool for evaluating the integrity of the calculated charges. Based on the SVD analysis for a selected group of molecules, we note the following:

- Increasing the molecular size reduces the fraction of charges which can be validly assigned.
- Increasing the point density of the fit has little or no influence on the rank of the problem.
- The symmetry problem in CHELP is due to statistical problems with the data and, contrary to common wisdom, is not entirely a function of the point density or point selection algorithm. In other words, there is generally no advantage to using CHELPG in place of CHELP. Both suffer from the ill conditioning of the matrix.

We also note that improvement in the rank can be achieved by selecting points closer to the molecular surface. Basis set has, as expected, no effect on the number of charges that can be assigned. Finally, we show that the SVD rank estimate can be used to generate improved sets of potential-derived charges.

Acknowledgments

M. M. F. would like to thank Professor V. J. Donnay for technical assistance during the preparation of the manuscript and Professor R. J. Woods for helpful discussions of the material.

References

1. See, for example, (a) H. A. Carlson, T. B. Nguyen, M. Orozco, and W. L. Jorgensen, *J. Comp. Chem.*, **14,** 1240 (1993); (b) P. Cieplak and P. A. Kollman, ibid., **12,** 1232 (1991); (c) K. Merz, ibid., **13,** 749 (1993); (d) J. J. Urban and G. R. Famini, ibid., **14,** 353 (1993); (e) listings of citations to other examples have appeared from time to time on the computational chemistry list (chemistry@osc.edu); see W. D. Cornell, September 26, 1994 (archives available via anonymous FTP at kekule.osc.edu).
2. For some recent reviews of methods for the assignment of charges to atoms in molecules, see (a) S. Bachrach, In *Reviews in Computational Chemistry*, Vol. 5, K. B. Lipkowitz and D. B. Boyd, Eds., VCH Publishers, New York, 1994, p. 171; (b) K. Wiberg and P. Rablen, *J. Comp. Chem.*, **14,** 1504 (1993).
3. R. S. Mulliken, *J. Chem. Phys.*, **23,** 1833 (1955).
4. For recent reviews, see R. F. W. Bader, *Chem. Rev.*, **91,** 893 (1991); R. F. W. Bader, P. L. A. Popelier, and T. A. Keith, *Angew. Chem., Intl. Ed. Engl.*, **33,** 620 (1994).
5. A. E. Reed, L. A. Curtiss, and F. Weinhold, *Chem. Rev.*, **88,** 899 (1988).
6. (a) S. R. Cox and D. E. Williams, *J. Comp. Chem.*, **2,** 304 (1981); (b) U. C. Singh and P. A. Kollman, ibid., **5,** 129 (1984); (c) L. E. Chirlian and M. M. Francl, ibid., **8,** 894 (1987); (d) C. M. Breneman and K. B. Wiberg, ibid., **11,** 361 (1990); (e) B. H. Besler, K. M. Merz, and P. A. Kollman, ibid., **11,** 431 (1990).
7. One such algorithm uses data from X-ray diffraction experiments to compute a molecular electrostatic potential to which charges are fit. See Z. Su, *J. Comp. Chem.*, **14,** 1036 (1993).
8. (a) T. R. Stouch and D. E. Williams, *J. Comp. Chem.*, **13,** 622 (1992); (b) T. R. Stouch and D. E. Williams, *J. Comp. Chem.*, **14,** 858 (1993).
9. R. J. Woods, M. Khalil, W. Pell, S. H. Moffat, and V. H. Smith, Jr., *J. Comp. Chem.*, **11,** 297 (1990).
10. R. Woods, personal communication and the computational chemistry list (chemistry@osc.edu), September 1, 1992 (archives available via anonymous FTP at kekule.osc.edu).
11. See, for example, B. S. Gottfried and J. Weisman, *Introduction to Optimization Theory*, Prentice Hall, Englewood Cliffs, NJ, 1973.
12. See, for example, (a) G. H. Golub and C. F. van Loan, *Matrix Computations*, 2nd ed., Baltimore, Johns Hopkins University Press, 1989; (b) C. L. Lawson and R. J. Hanson, *Solving Least Squares Problems*, Prentice Hall, Englewood Cliffs, NJ, 1974; (c) W. H. Press, B. P. Flannery, S. A. Teukolsky, and W. T. Vetterling, *Numerical Recipes*, Cambridge University Press, Cambridge, UK, 1986.
13. C. I. Bayly, P. Cieplak, W. D. Cornell, and P. A. Kollman, *J. Phys. Chem.*, **97,** 10269 (1993).
14. L. E. Chirlian and M. M. Francl, *QCPE*, **7,** 39 (1987).
15. M. M. Francl, R. F. Hout, Jr., and W. J. Hehre, *J. Am. Chem. Soc.*, **106,** 563 (1984).
16. M. J. Frisch, G. W. Trucks, M. Head-Gordon, P. M. W. Gill, M. W. Wong, J. B. Foresman, B. G. Johnson, H. B. Schlegel, M. A. Robb, E. S. Replogle, R. Gomperts, J. L. Andres, K. Raghavachari, J. S. Binkley, C. Gonzalez, R. L. Martin, D. J. Fox, D. J. Defrees, J. Baker, J. J. P. Stewart, and J. A. Pople, Gaussian 92, revision C, Gaussian, Inc., Pittsburgh, PA, 1992.
17. SPARTAN, Wavefunction, Inc., Irvine, CA.
18. S. Abrahmsson and I. Pascher, *Acta Cryst.*, **21,** 79 (1966).
19. (a) First-row elements: J. S. Binkley, J. A. Pople, and W. J. Hehre, *J. Am. Chem. Soc.*, **102,** 939 (1980); (b) second-row elements: W. J. Pietro, M. M. Francl, W. J. Hehre, D. J. Defrees, J. A. Pople, and J. S. Binkley, *J. Am. Chem. Soc.*, **104,** 5039 (1982).
20. R. Krishnan, M. J. Frisch, and J. A. Pople, *J. Chem. Phys.*, **72,** 650 (1980); (b) M. J. Frisch, J. S. Binkley, and J. A. Pople, *J. Chem. Phys.*, **80,** 3265 (1984).
21. See M. M. Francl, *J. Phys. Chem.*, **89,** 428 (1985) and C. H. Douglass, Jr., D. A. Weil, P. A. Charlier, R. A. Eades, D. G. Truhlar, and D. A. Dixon, In *Chemical Applications of Atomic and Molecular Potentials*, P. Politzer and D. G. Truhlar, Eds., Plenum Press, New York, 1981.